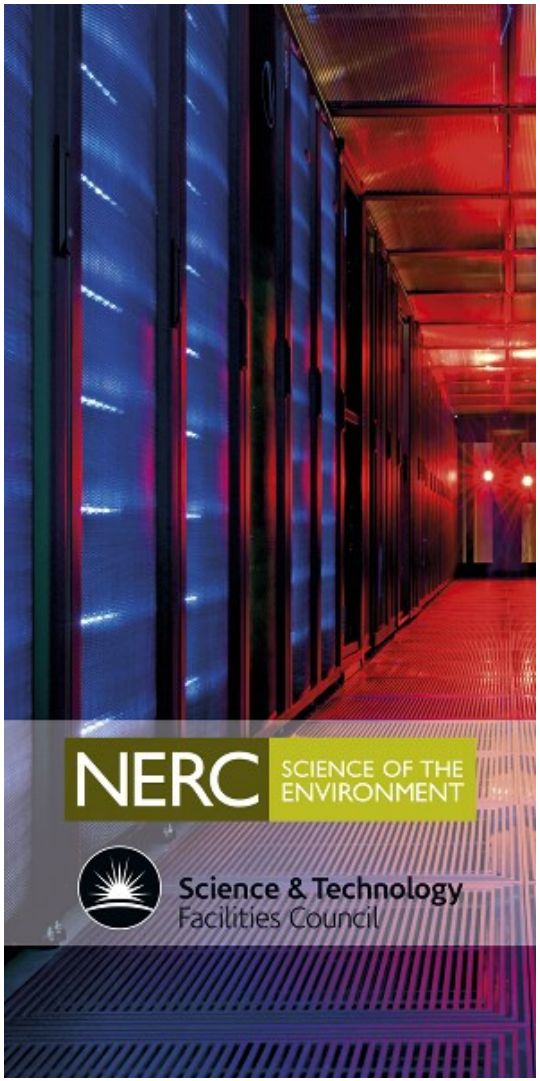
Science & Technology Facilities Council
Rutherford Appleton Laboratory



JASMIN
Petascale storage and terabit networking for environmental science

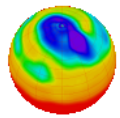
Matt Pritchard
Centre for Environmental Data Archival
RAL Space

Jonathan Churchill
Scientific Computing Department

STFC Rutherford Appleton Laboratory

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

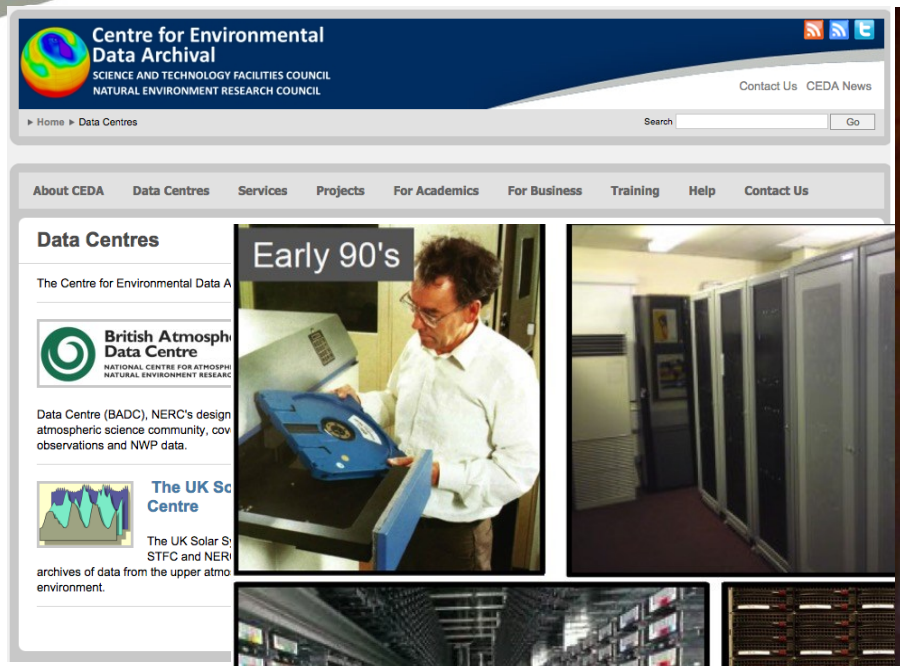
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL



What is JAMSIN?

- What is it?
  - 16 Petabytes high-performance disk
  - 4000 computing cores
    - (HPC, Virtualisation)
  - High-performance network design
  - Private clouds for virtual organisations
- For Whom?
  - **Entire NERC community**
  - Met Office
  - European agencies
  - Industry partners
- How?

Context


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Contact Us CEDA News
Home Data Centres
Search Go
About CEDA Data Centres Services Projects For Academics For Business Training Help Contact Us
Data Centres
The Centre for Environmental Data A
British Atmosph Data Centre
Data Centre (BADC), NERC's design
atmospheric science community, cov
observations and NWP data.
The UK Sc Centre
The UK Solar S
STFC and NER
archives of data from the upper atmo
environment.

Early 90's


JASMINE


2014

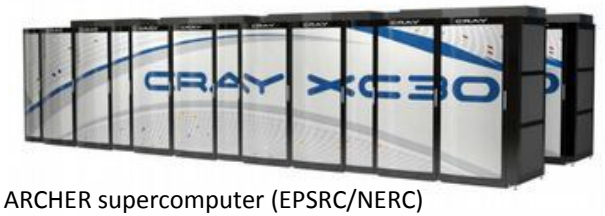
JASMIN: the missing piece

- Urgency to provide better environmental predictions
- Need for higher-resolution models
- HPC to perform the computation
- Huge increase in observational capability/capacity

But…

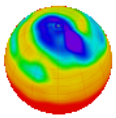
- Massive storage requirement: observational data transfer, storage, processing
- Massive raw data output from prediction models
- Huge requirement to process raw model output into usable predictions (graphics/post-processing)

Hence JASMIN…



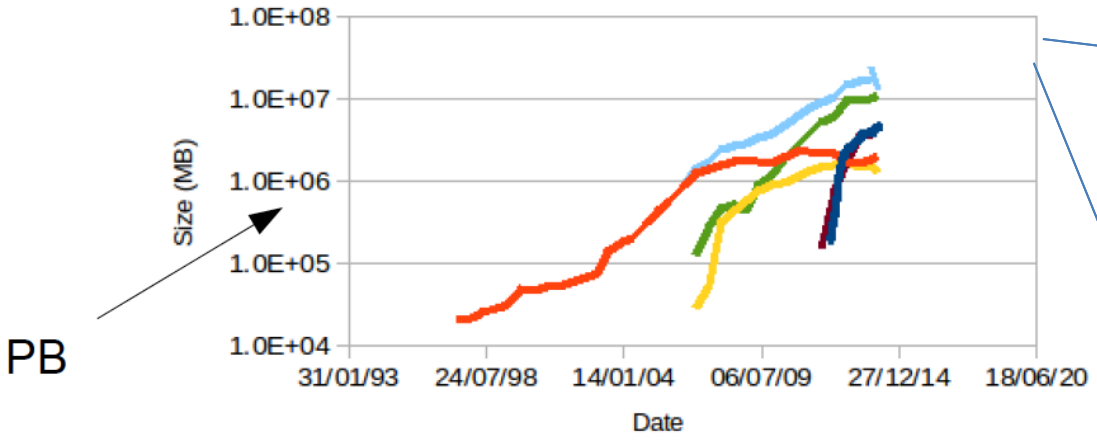
ARCHER supercomputer (EPSRC/NERC)

JAMSIN (STFC/Stephen Kill)



Data growth

(Credit: Folkes, Churchill)



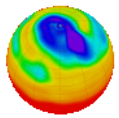
30-300 Pb
?CMIP6 Climate models

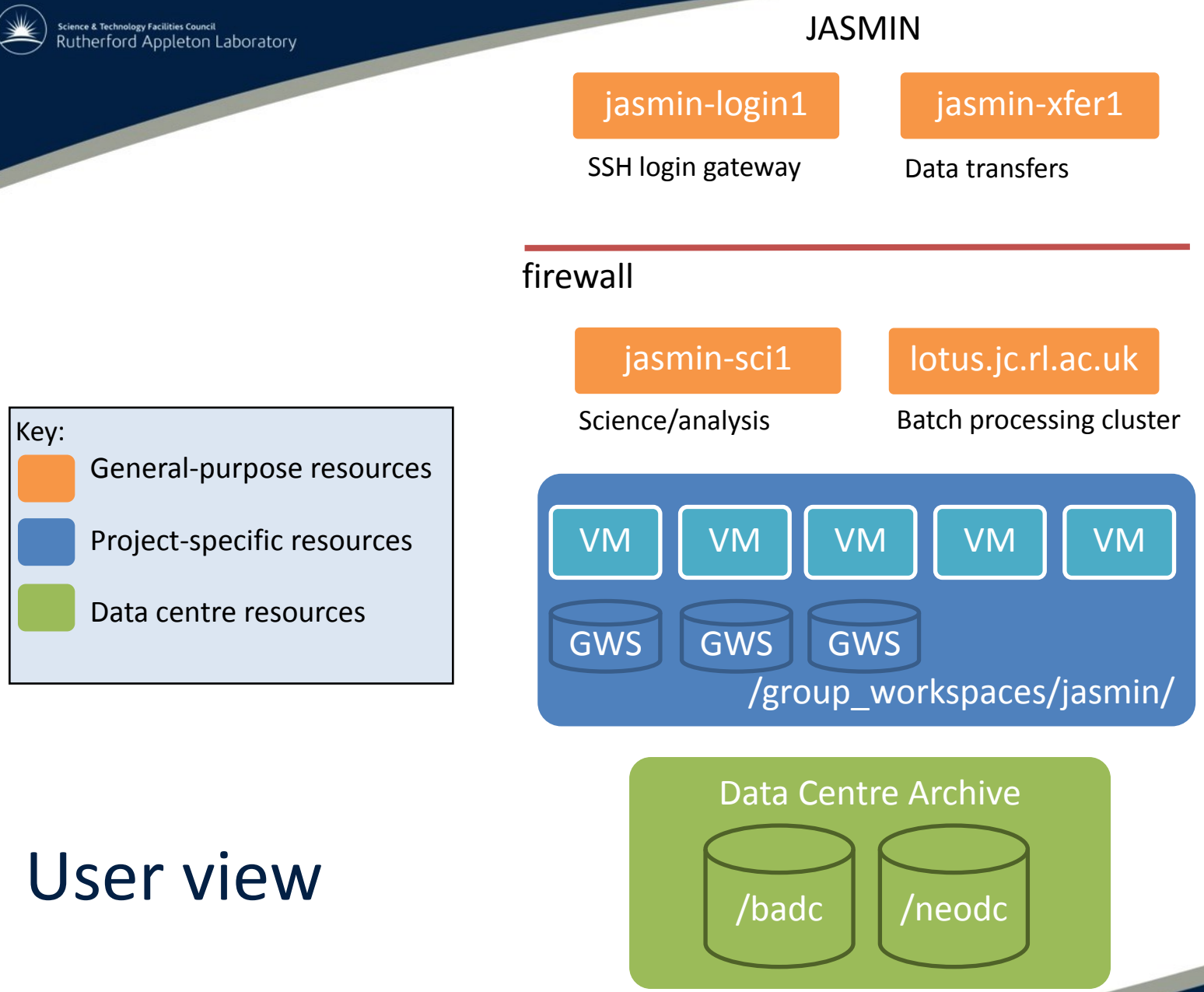
30-85 Pb (unique data) projection for JASMIN

Light blue = total of all tape at STFC
Green = Large Hadron Collider (LHC) Tier 1 data on tape
Dark blue = data on **disk** in JASMIN

Data growth on JASMIN has been limited by:

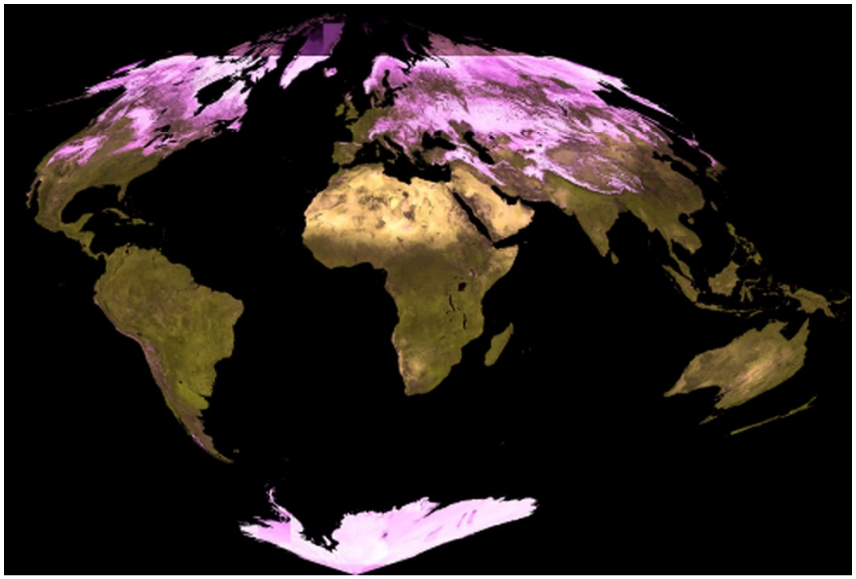
Not enough disk	(now fixed …for a while)
Not enough local compute	(now fixed …for a while)
Not enough inbound bandwidth	(now fixed …for a while)




JASMIN
jasmin-login1
jasmin-xfer1
SSH login gateway
Data transfers
firewall
jasmin-sci1
lotus.jc.rl.ac.uk
Science/analysis
Batch processing cluster
Key:
General-purpose resources
Project-specific resources
Data centre resources
VM
VM
VM
VM
VM
GWS
GWS
GWS
/group_workspaces/jasmin/
Data Centre Archive
/badc
/neodc

User view

Success stories (1)



- QA4ECV
  - Re-processed MODIS Prior in 3 days on JASMIN-LOTUS, 81x faster than original process on 8-core blade
  - Throw hardware* at the problem

*Right type of hardware

Success stories (2)

- Test feasible resolutions for global climate models
- 250 Tb in 1 year generated on PRACE supercomputing facility in Germany (HERMIT). 400 Tb total.
- Network transfer to JASMIN
- Analysed by Met Office scientists as soon as available
- Deployment of VMs running custom scientific software, co-located with data
- Outputs migrated to long term archive (BADC)

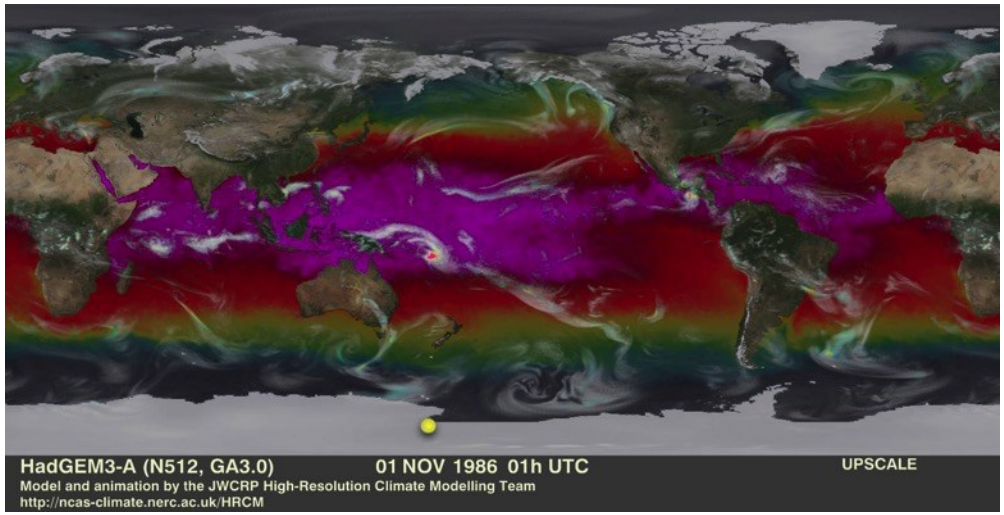


Image: P-L Vidale & R. Schiemann, NCAS

Mizielinksi et al (Geoscientific Model Development, 2013)
"High resolution global climate modelling; the UPSCALE project, a large simulation campaign"

Coming soon

GOV.UK

Search

Departments Worldwide How government works Get involved
Policies Publications Consultations Statistics Announcements

News story

Europe's Earth observation programme maximised by UK data hub

From: UK Space Agency
First published: 24 March 2015
Part of: Science and innovation and UK economy

The UK is to host a world-class data facility, giving scientists full access to Earth observation data from Europe's Copernicus programme.

Signature of ESA/UK Collaborative Ground Segment Cooperation agreement.
Credit: ESA–N. Imbert-Vier, 2015.

esa

Copernicus Space Component
Ground Segment and Operations
Concept

Phase 2/3 expansion

- 2013 NERC Big Data capital investment
  - Wider scope: support projects from new communities, e.g.
    - EOS Cloud
      - Environmental 'omics. Cloud BioLinux platform
    - Geohazards
      - Batch compute of Sentinel-1a SAR for large-scale, hi-res Earth surface deformation measurement

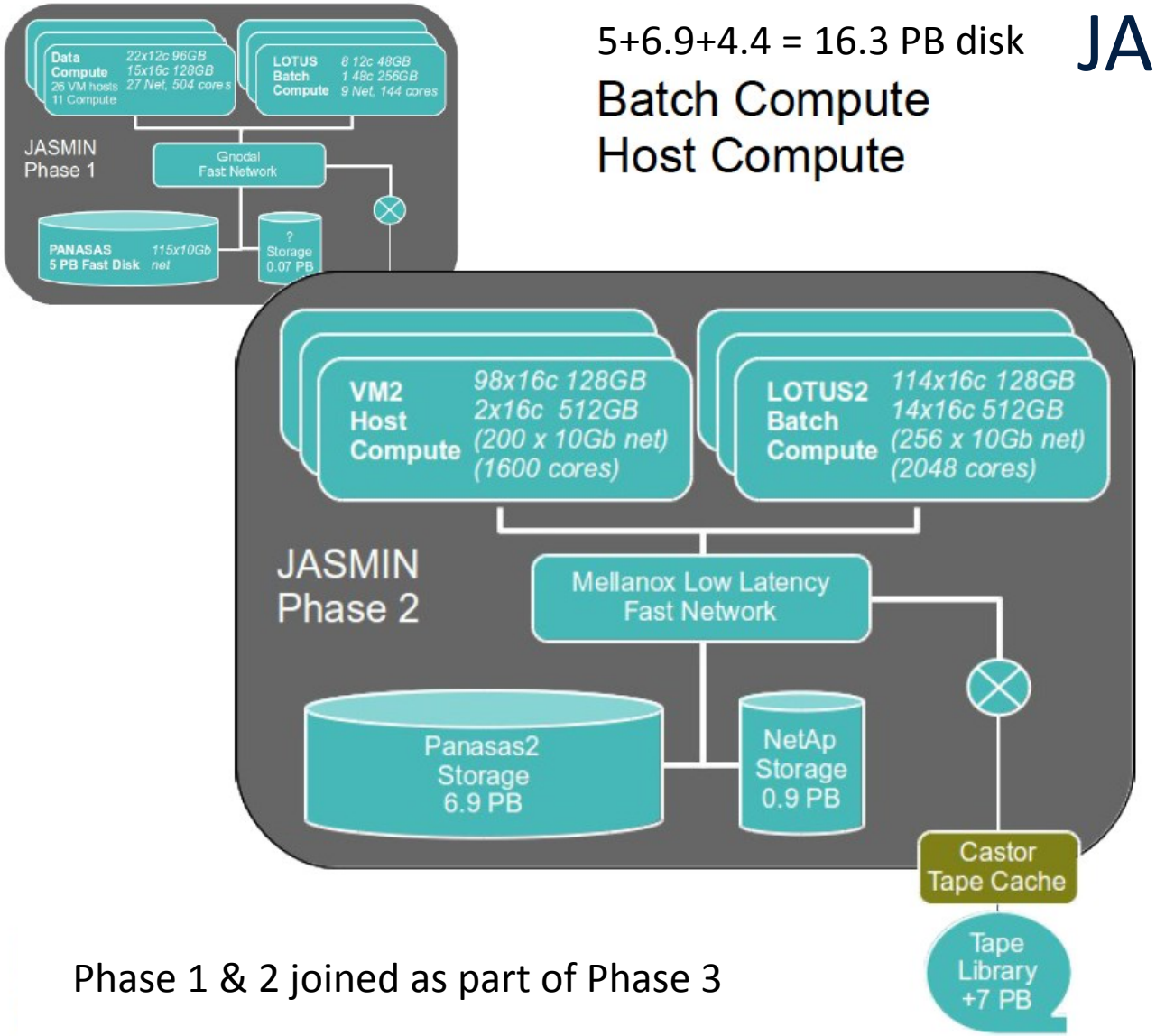
Sentinel-1a (ESA)

	Phase 2 by March 2014	Phase 3 by March 2015
JASMIN hard upgrade	+7 Petabytes disk +6 Petabytes tape +3000 compute cores network enhancement	+o(2) Petabytes disk +o(800) compute cores network enhancement
JASMIN soft upgrade	Virtualisation software Scientific analysis software Cloud management software Dataset construction Documentation	

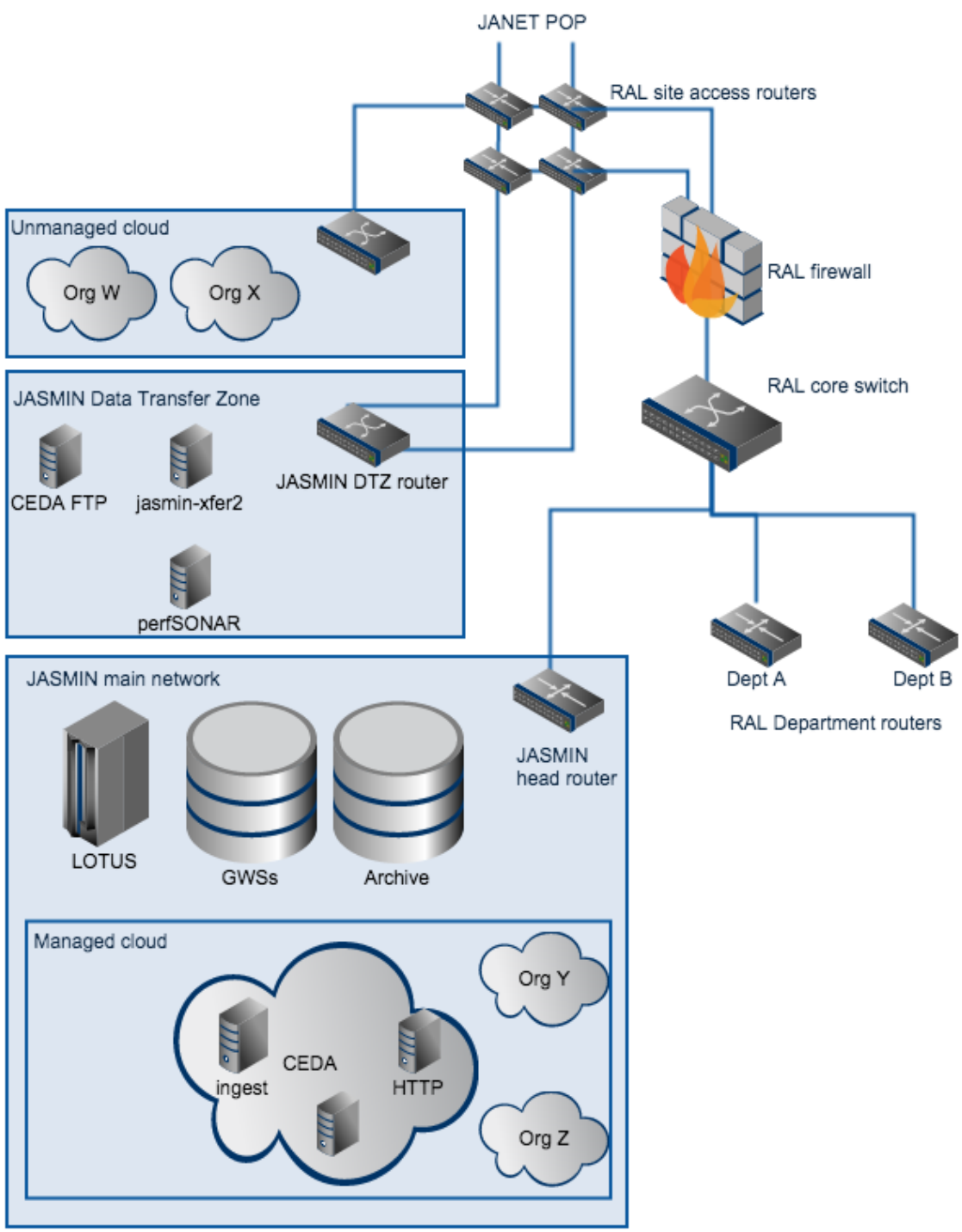
JASMIN Now

5+6.9+4.4 = 16.3 PB disk

Batch Compute
Host Compute

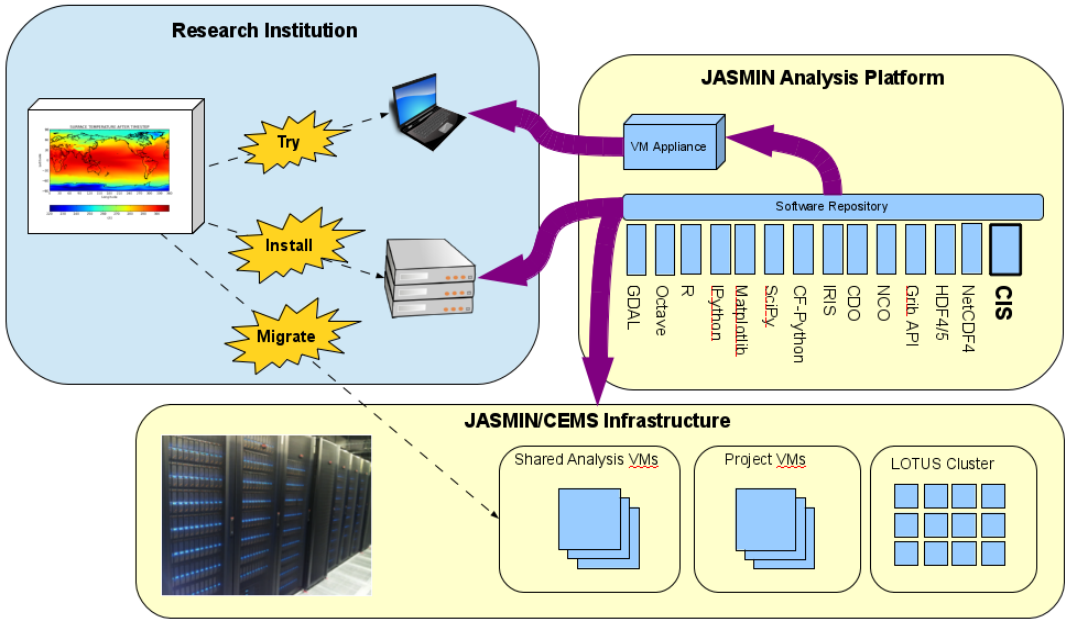


Phase 1 & 2 joined as part of Phase 3


JANET POP
RAL site access routers
Unmanaged cloud
Org W
Org X
RAL firewall
JASMIN Data Transfer Zone
CEDA FTP
jasmin-xfer2
JASMIN DTZ router
RAL core switch
perfSONAR
Dept A
Dept B
RAL Department routers
JASMIN main network
JASMIN
head router
LOTUS
GWSs
Archive
Managed cloud
Org Y
ingest
CEDA
HTTP
Org Z

JASMIN Analysis Platform

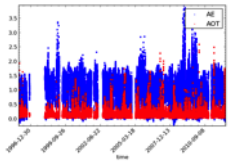
- Software stack for scientific analysis on JASMIN
  - Common packages for climate science, geospatial analysis
  - **CIS** developed as part of JASMIN
  - Deployed on JASMIN
    - Shared VMs
    - Sci VM template
    - LOTUS



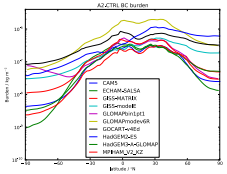
Community Intercomparison Suite (CIS)

CIS = Component of JAP

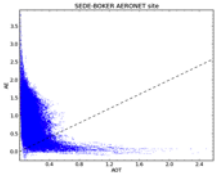
Time-series



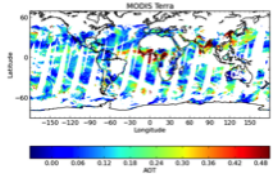
Line plots

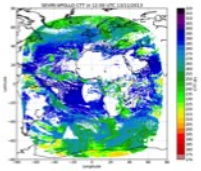


Scatter plots

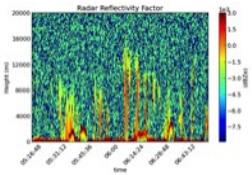


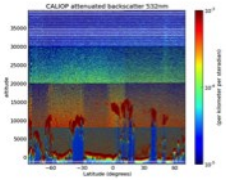
Global plots



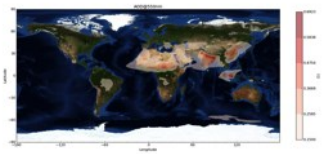


Curtain plots

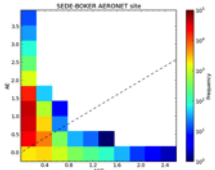




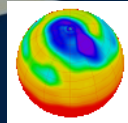
Overlay plots



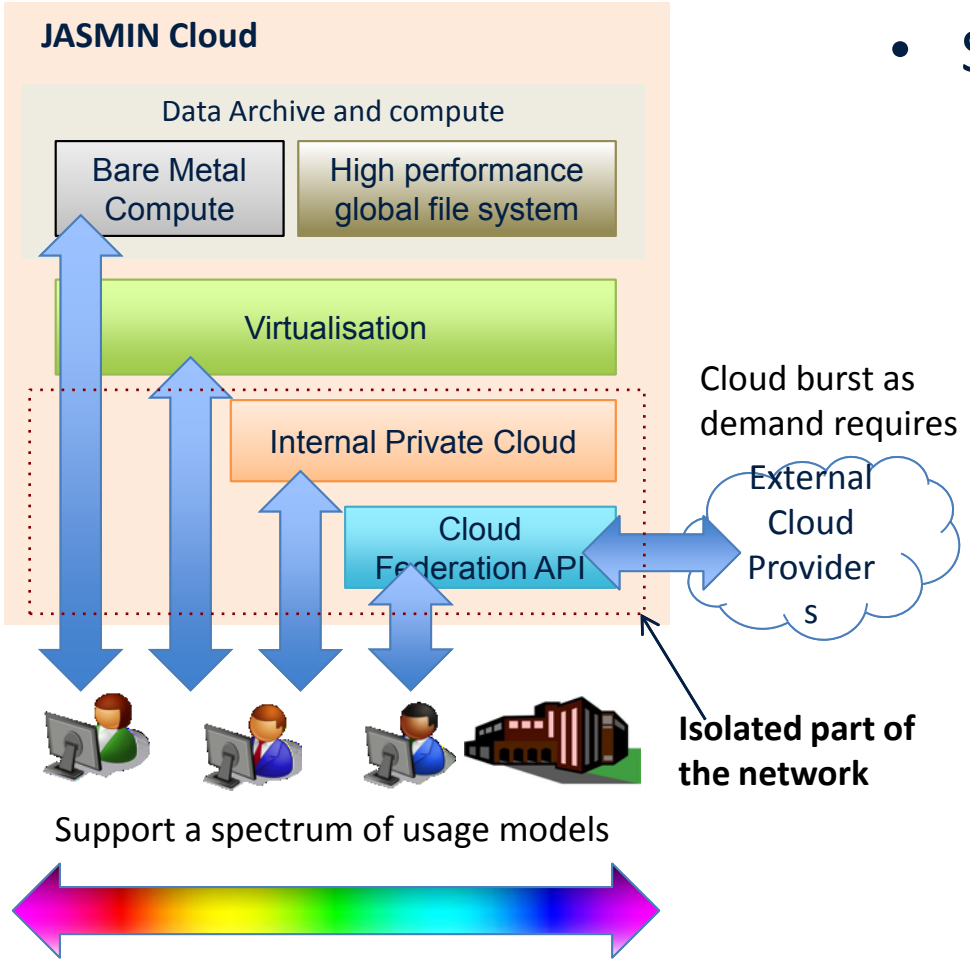
Histograms



Dataset	Format
AERONET	Text
MODIS	HDF
CALIOP	HDF
CloudSAT	HDF
AMSRE	HDF
TRMM	HDF
CCI aerosol & cloud	NetCDF
SEVIRI	NetCDF
Flight campaign data	RAF
Models	NetCDF



Vision for JASMIN 2 (Applying lessons from JASMIN 1)

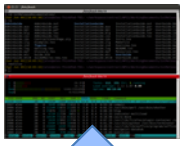


- Some key features
  - Nodes are general purpose: boot as bare metal or hypervisors
  - Use cloud tenancy model to make Virtual Organisations
  - Networking: made isolated network inside JASMIN to give users greater freedom: full IaaS, root access to hosts …

JASMIN Cloud Architecture

IPython Notebook VM could access cluster through Python API

ssh via public IP



JASMIN Cloud Management Interfaces

CloudBioLinux Desktop

ASMIN Internal Network

External Network *inside* JASMIN

Managed Cloud - PaaS, SaaS

Science Analysis VM

Storage

Project1-org

Login VM

Science Analysis VM

Storage

Project2-org

Compute Cluster VM

CloudBioLinux Fat Node

File Server VM

Storage

eos-cloud-org

Direct access to batch processing cluster

Direct File System Access

Standard Remote Access Protocols – ftp, http, …

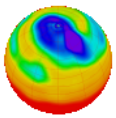
Lotus Batch Compute

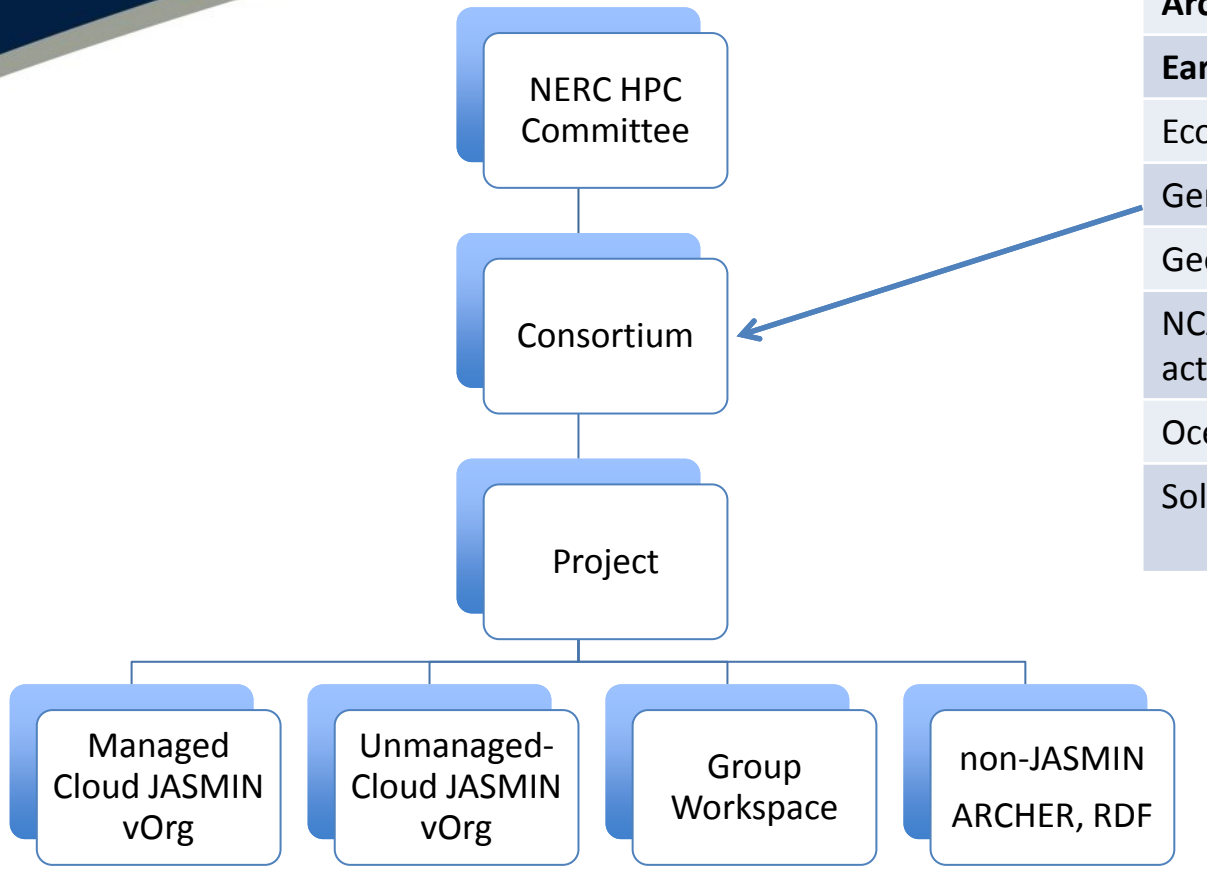
Data Centre Archive

Unmanaged Cloud – IaaS, PaaS, SaaS

First cloud tenants

- EOS Cloud
  - Environmental bio-informatics
- MAJIC
  - Interface to land-surface model
- NERC Environmental Work Bench
  - Cloud-based tools for scientific workflows





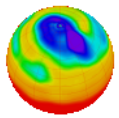
Consortia
Atmospheric & Polar Science
Archive
Earth Observation & Climate Services
Ecology & Hydrology
Genomics
Geology
NCAS CMS Director cross-cutting activities
Oceanography & Shelf Seas
Solid Earth & Mineral Physics

a Tb disk
b GB RAM
c vCPU

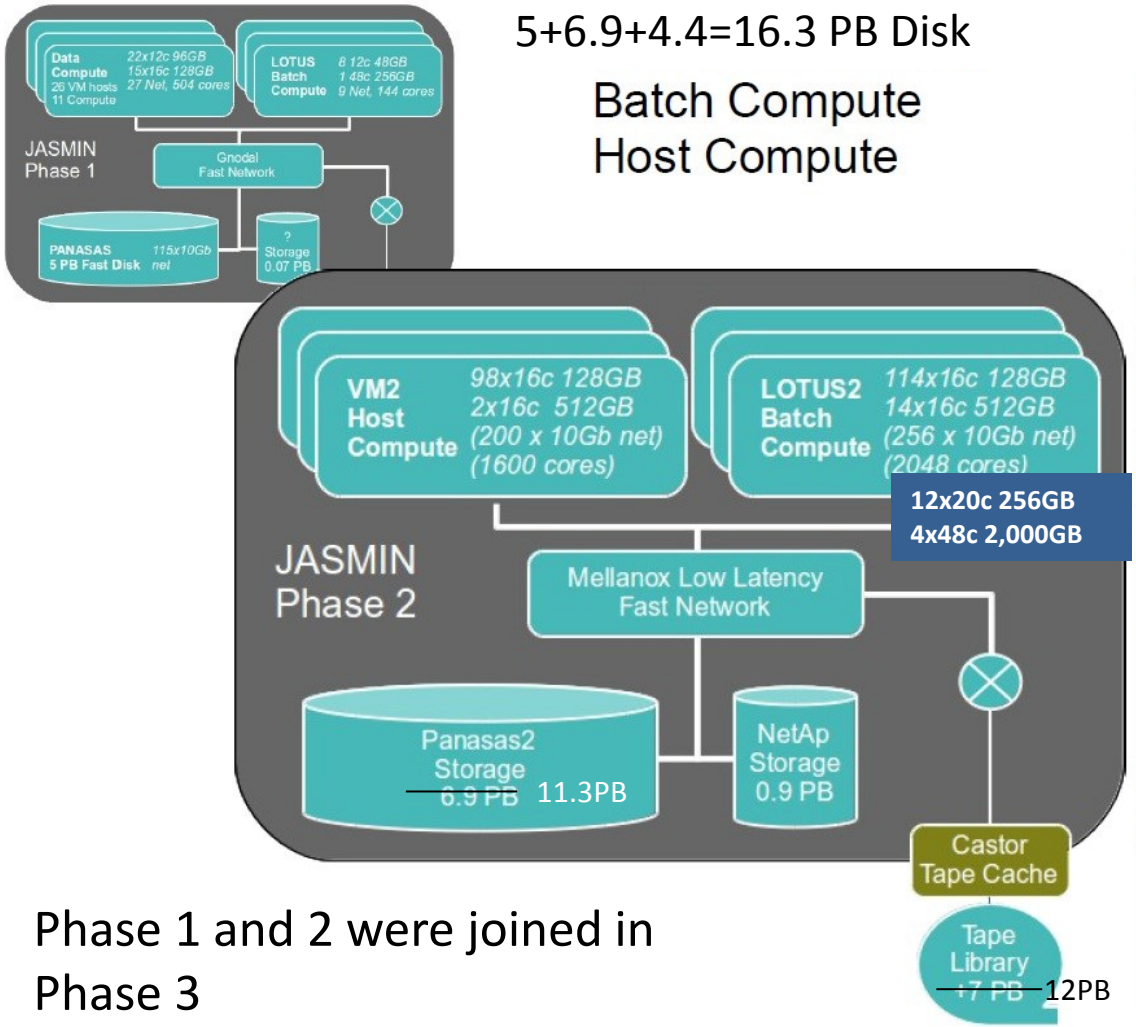
d Tb disk
e Gb RAM
f vCPU

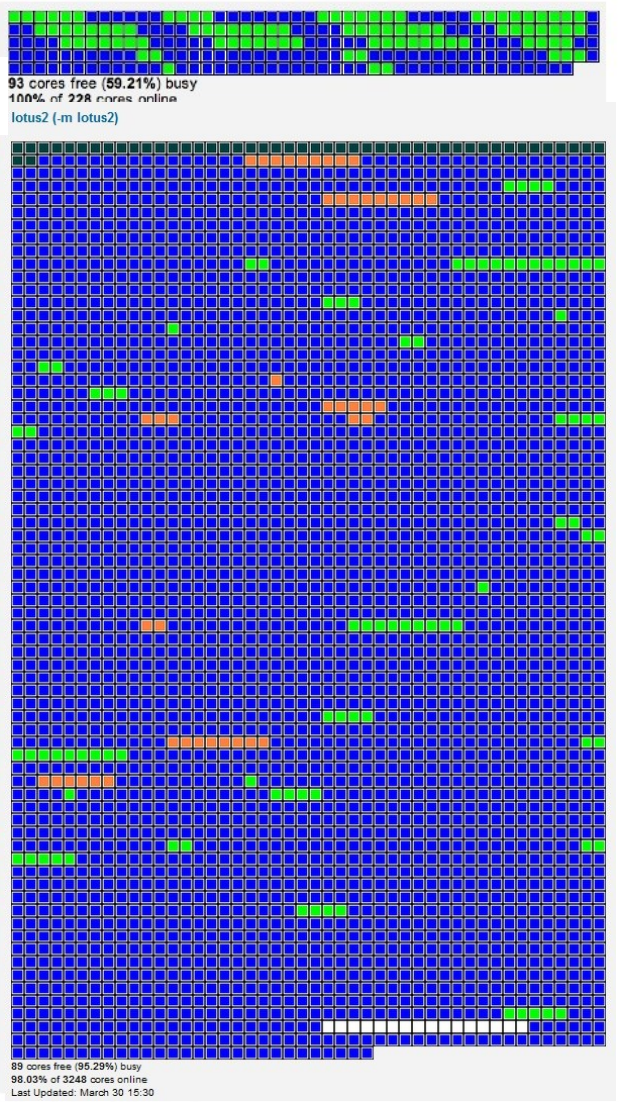
x Tb disk

x Tb RDF disk
y AU ARCHER



JASMIN Now

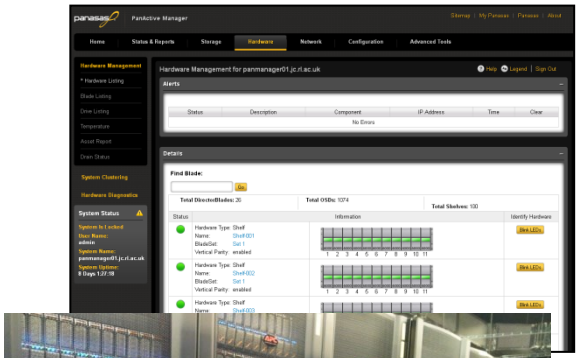


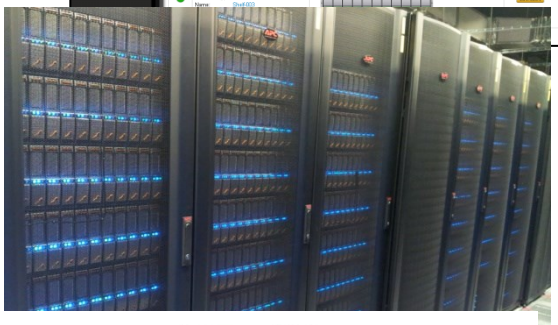


Phase 1 and 2 were joined in Phase 3

Panasas Storage

- Parallel file system (cf Lustre, GPFS, pNFS etc)
  - Single Namespace
  - 140GB/sec benchmarked (95 shelves PAS14)
  - Access via PanFS client/NFS/CIFS
  - Posix filesystem out of the box.
- Mounted on Physical machines and VMs
- 103 shelves PAS11
  + 101 shelves PAS14
  + 40 Shelves PAS16
  - Each shelf connected at 10Gb (20Gb PAS14)
  - 2,684 'Blades'
  - Largest single realm & single site in the world
- One Management Console
- TCO: *Big Capital, Small Recurrent*
  but JASMIN2 £/TB < GPFS/Lustre offerings





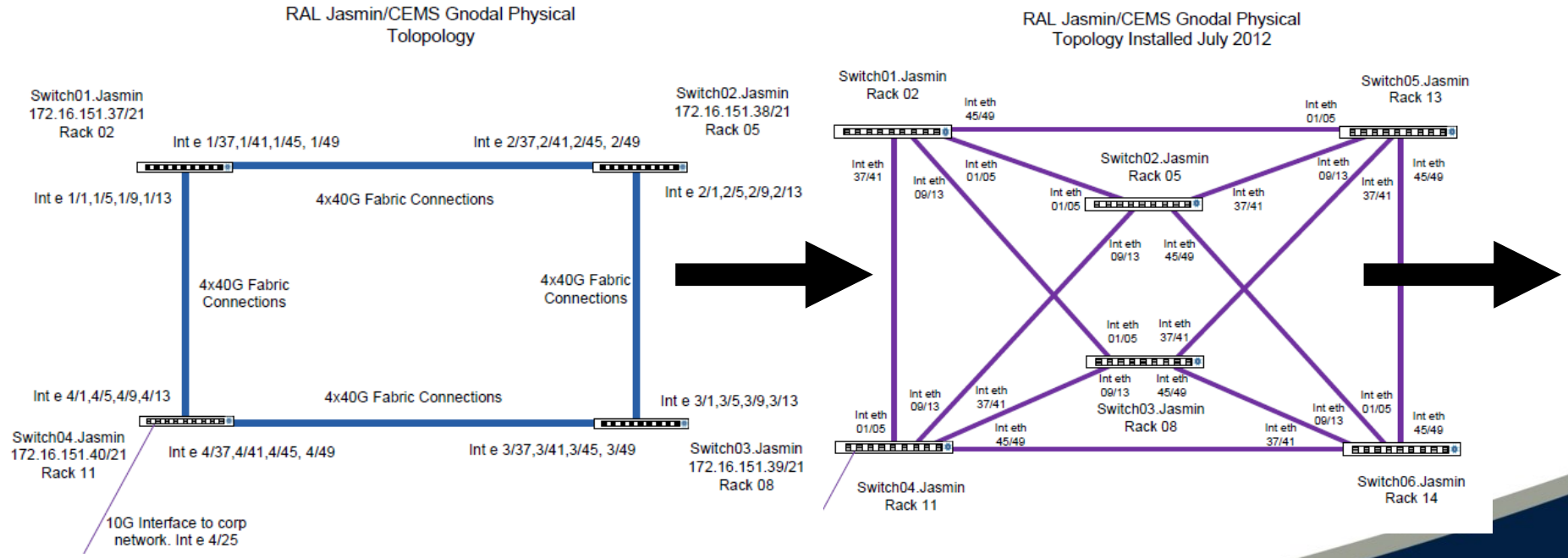
Three year growth pains

172.16.144.0/21 = 2,000 IPs

130.246.136.0/21

Flat Overlaid L2

160->240 Ports @ 10Gb



Science & Technology Facilities Council
Rutherford Appleton Laboratory

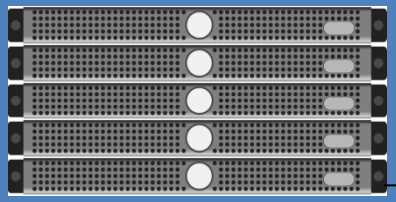
Overview

£12.5M,38 Racks, 850Amps, 25 tonnes,

3Terabit/s bandwidth

4 x VMware Clusters

vJASMIN
156 cores, 1.2TB

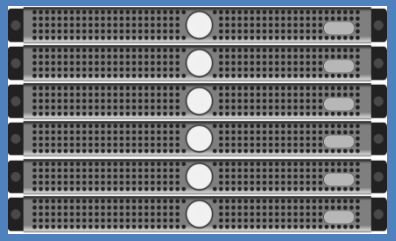


40x 10Gb

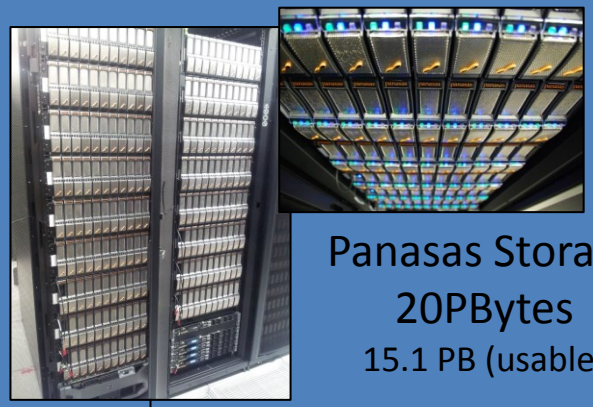
32x 10Gb

NetApp+Dell 1010TB +
(VM VMDK images)

vCloud
208-1648 cores,
1.5TB – 12.8TB



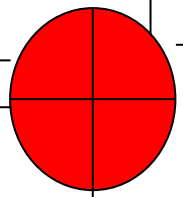
40x 10Gb



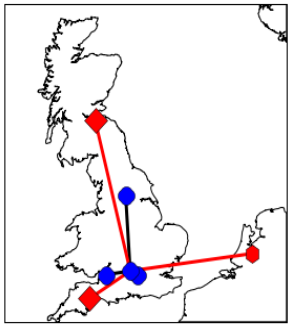
Panasas Storage
20PBytes
15.1 PB (usable)

385 x 10Gb

468 x 10Gb

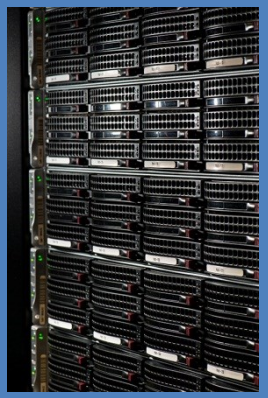


A network :
1,100 Ports @ 10GbE



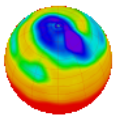
LightPaths @ 1&2Gb/s and 10Gb/s:
Leeds, UKMO, Archer, CEMS-SpaceCat

Lotus HPC Cluster



MPI network
(10Gb low latency eth)

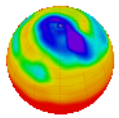
144-234 hosts, 2.2K-3.6K cores.
RHEL6, Platform LSF, MPI



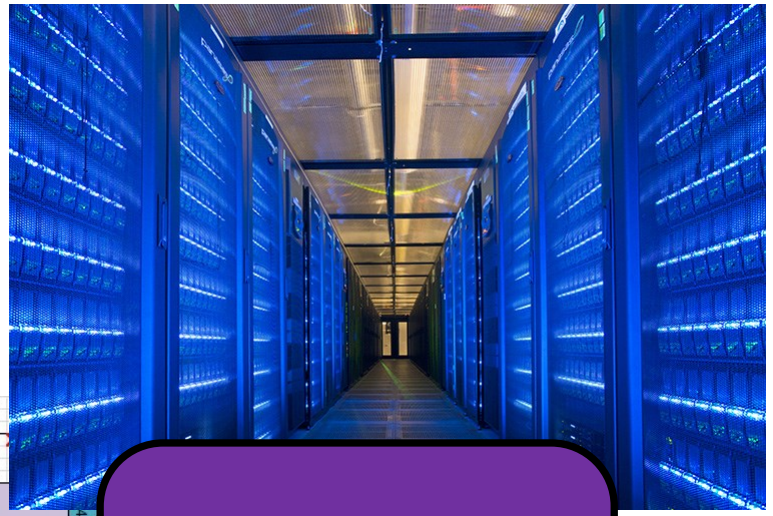
Network Design Criteria

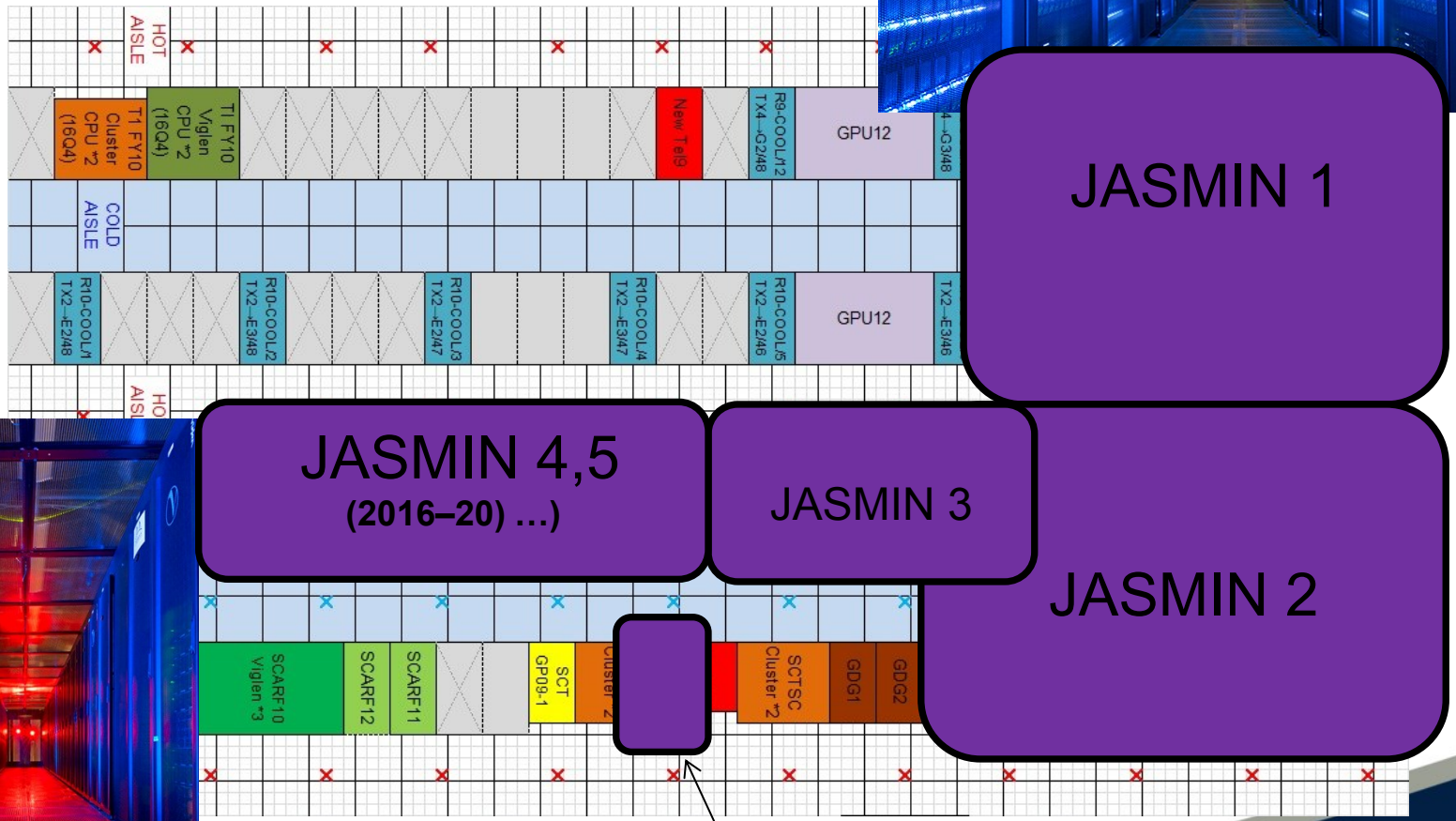
- Non-Blocking (No network contention)
- Low Latency (< 20uS MPI. Preferably < 10uS)
- Small latency spread.
- Converged (IP storage, SAN storage, Compute, MPI)
- 700-1100 Ports @ 10Gb
- Expansion to 1,600 ports and beyond wo forklift.
- Easy to manage and configure
- Cheap
- …… later on:

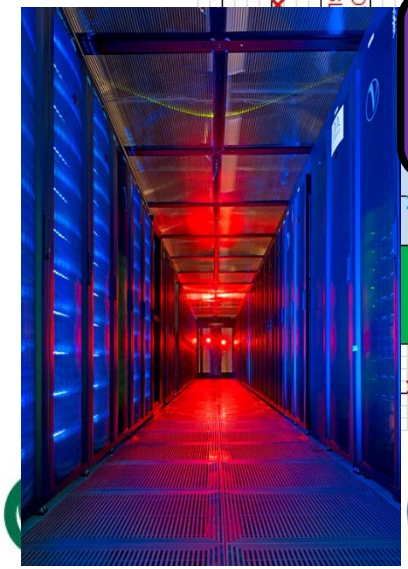
  Replaces JASMIN1 240 ports in place.



Floor Plan
Network distributed ~30m x ~20m

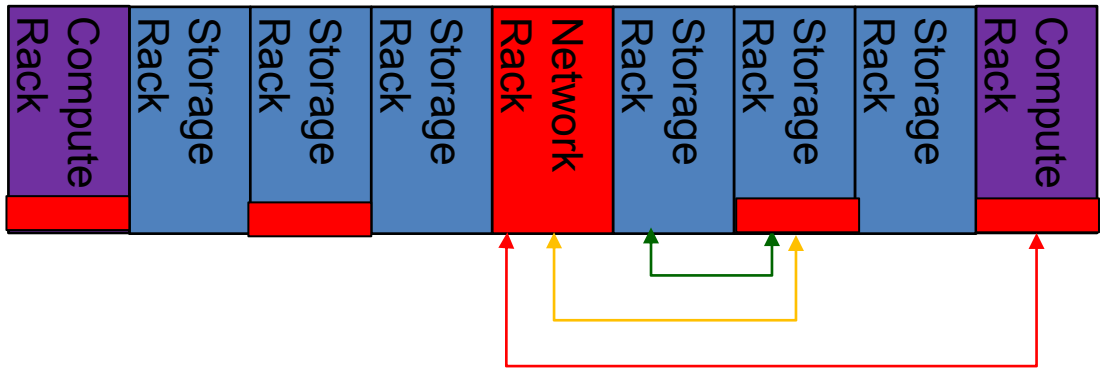






National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Cabling Costs: Central Chassis Switch + ToR



312 Twinax

<u>Fully Populated ToR</u>
6x S4810 ToR Switches
48x Active Optic QSFP

<u>1:1 Contention ToR</u>
20x S4810 ToR Switches
80x Active Optic QSFP

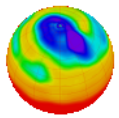
ToR e.g Force10 S4810P
48 x 10Gb SFP+
4x 40Gb QSFP+

1,000 Fibre Connections = £400-600K
JASMIN1+2 700-1100 10Gb Connections

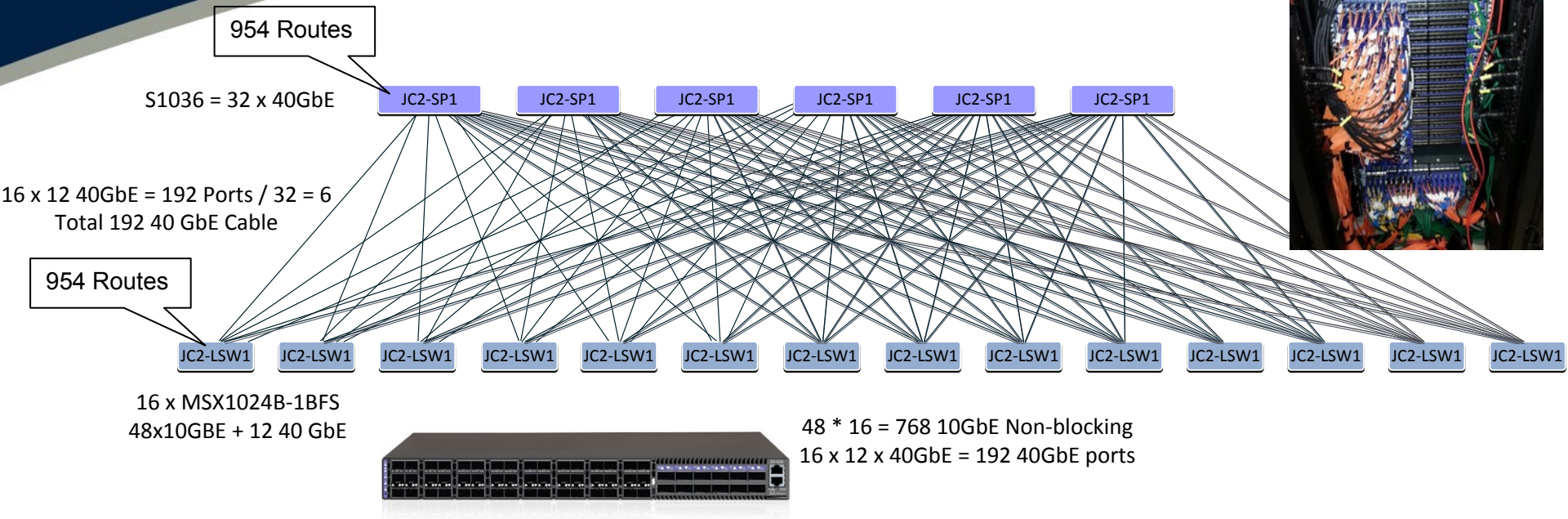
Lots of core 40Gb ports needed.
MLAG to 72 Ports ?....
Chassis switch ? … expansion/cost

ECMP Core

Spine
12 Switches - each 36 x 40GbE
40GbE
40GbE
1:1
32 Switches - each 48 x 10GbE
12x40
Leaf
48x10
10GbE
1:1

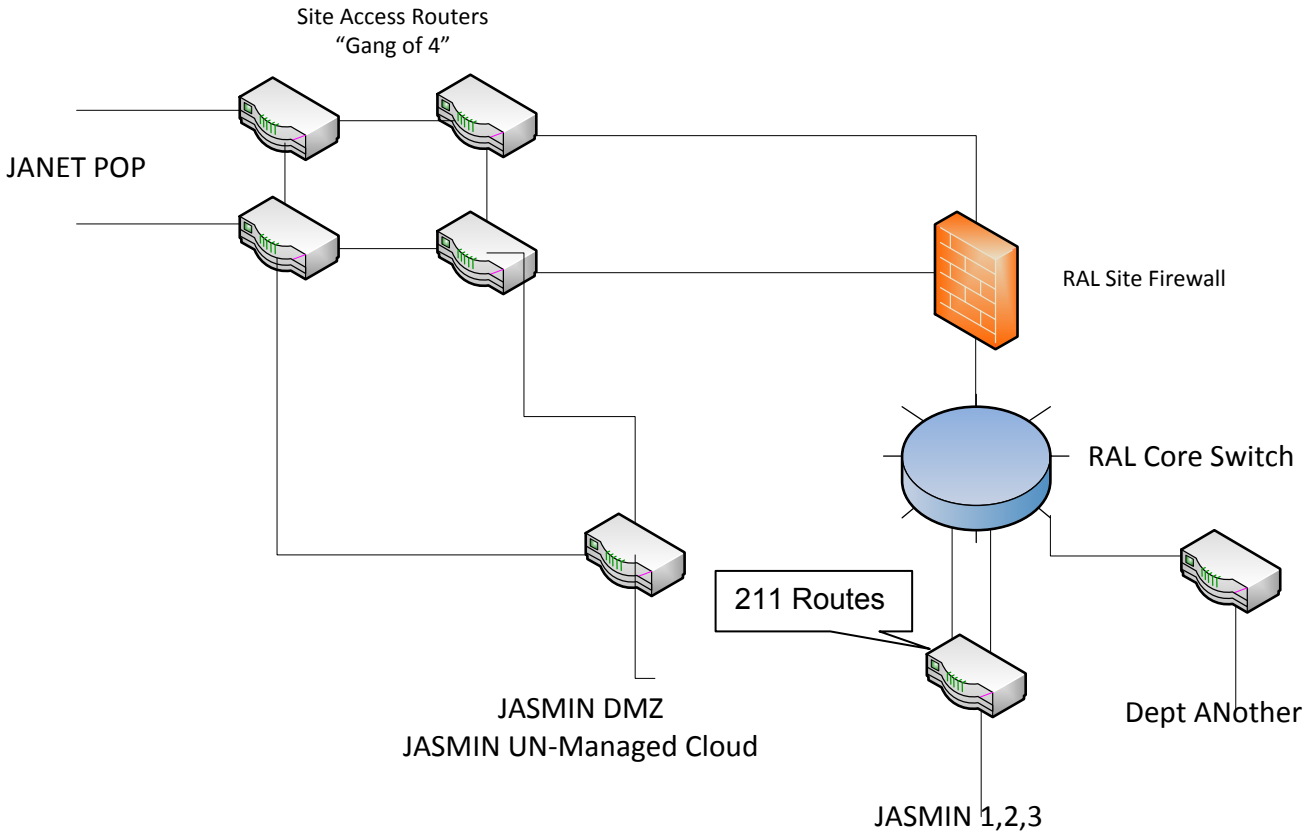


1,104 x 10GbE Ports CLOS L3 ECMP OSPF



- 768 Ports max. no expansion … so 12 spines
- Max 36 leaf switches :1,728 Ports @ 10GbE
- Non-Blocking. Zero Contention (48x10Gb = 12x 40Gb uplinks)
- Low Latency (250nS L3 / per switch/router). 7-10uS MPI
- Cheap ! ….. (ish)

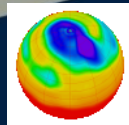
RAL Site Network (for comparison)



ECMP CLOS L3 Advantages

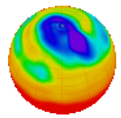
- Massive scale
- Low cost (Pay as you grow)
- High performance
- Low latency
- Standards based – supports multiple vendors
- Very small “blast radius” upon network failures
- Small isolated subnets
- Deterministic latency with a fixed spine and leaf

https://www.nanog.org/sites/default/files/monday.general.hanks.multistage.10.pdf

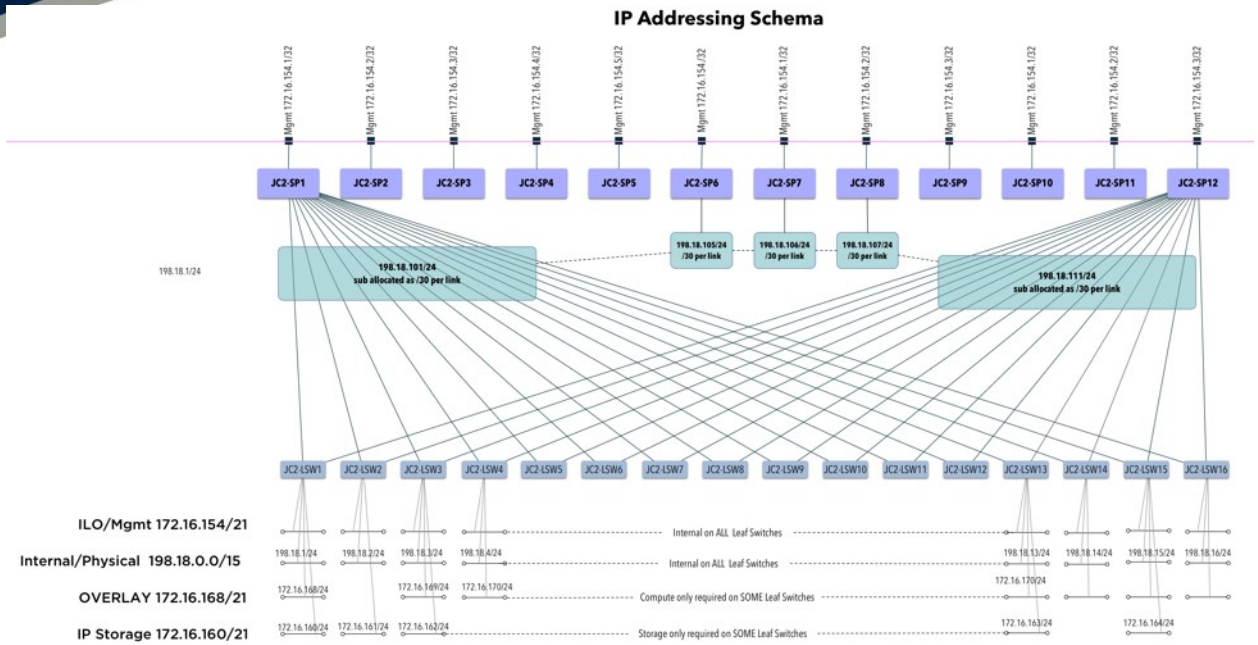


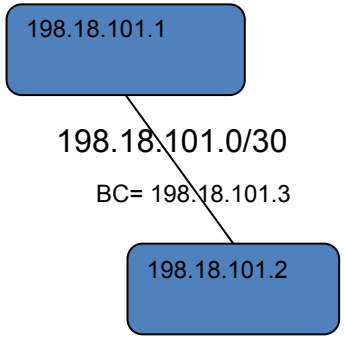
ECMP CLOS L3 Issues

- Managing scale:
  - #s of IPs, subnets, VLANs, Cables
  - Monitoring
- Routed L3 network:
  - Reqs dynamic OSPF routing (100's routes per switch)
  - No L2 between switches (VMware: SAN's, vMotion)
    - Reqs: DHCP Relay, VXLAN
  - Complex 'traceroute' seen by users.



IP and Subnet Management

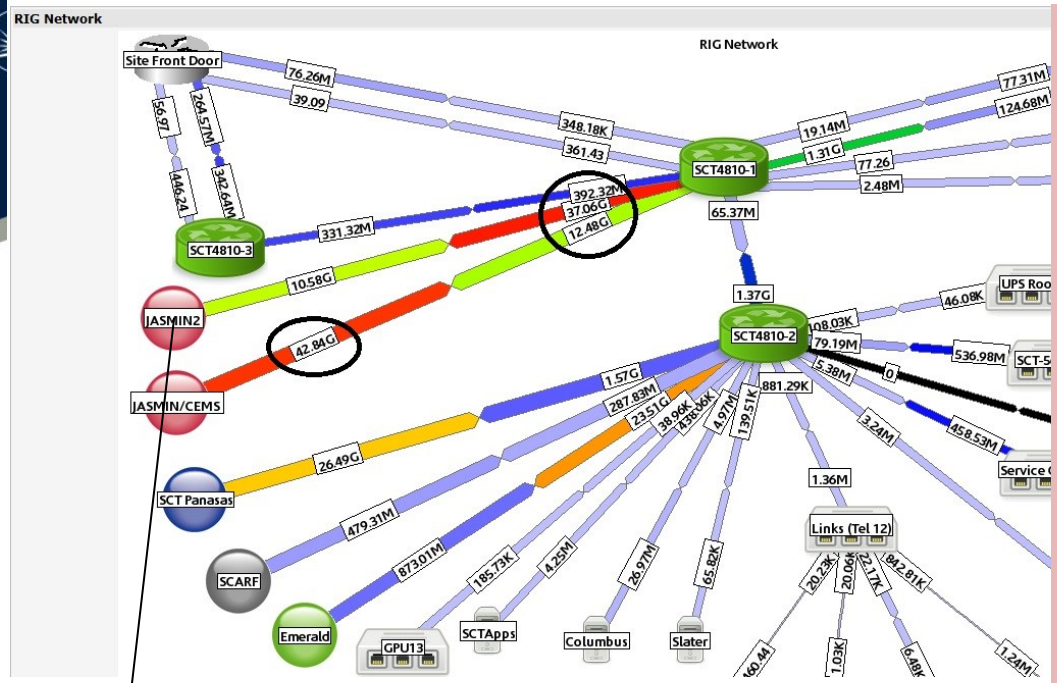


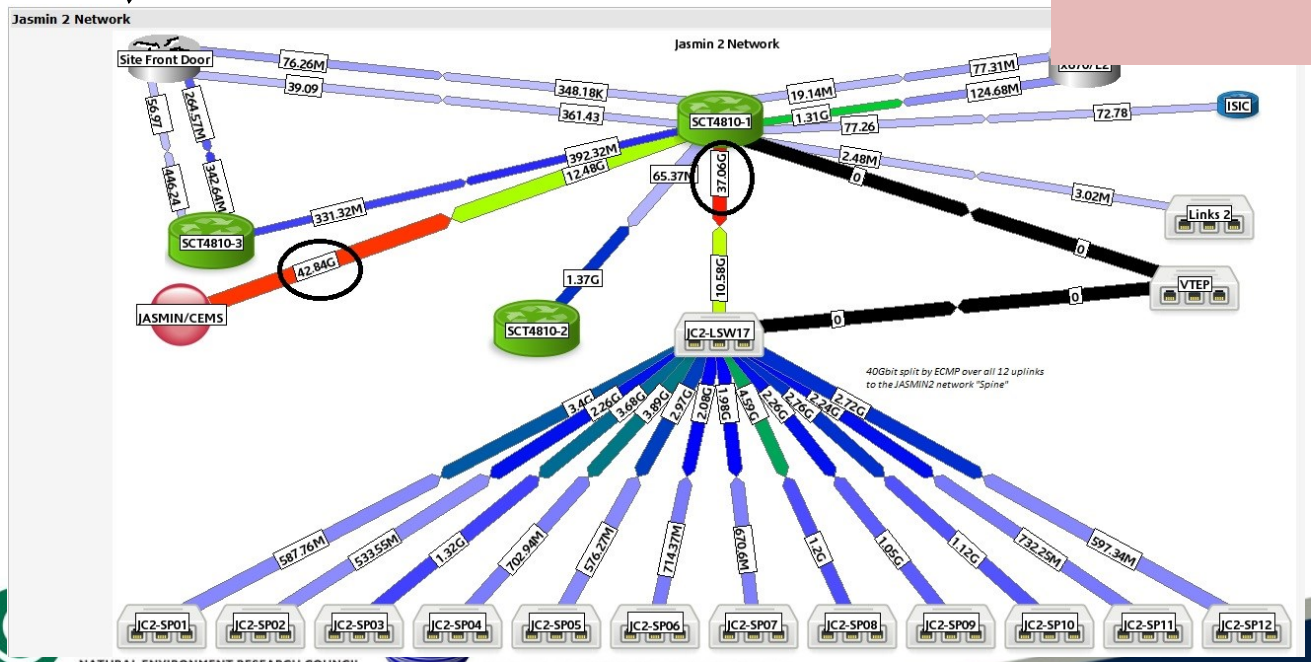


4 IPs / Cable

- 2x /21 Panasas Storage
- 4x /24 Internet Connects
- 55x /26 Fabric Subnets
- 264x /30 Inter switch links

- 514 VMs & Growing quickly
- 304 Servers, 850 Switches/PDUs etc
- 2,684 Storage Blades
- ~260 VLAN IDs
- 6,000 IPs & Growing





Monitoring / Visualisation

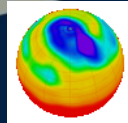
- Complex Cacti
  - >30 Fabric Switches
  - >50 Mgmt Switches
- 100's links to monitor
- Nagios bloat

root@host052:~

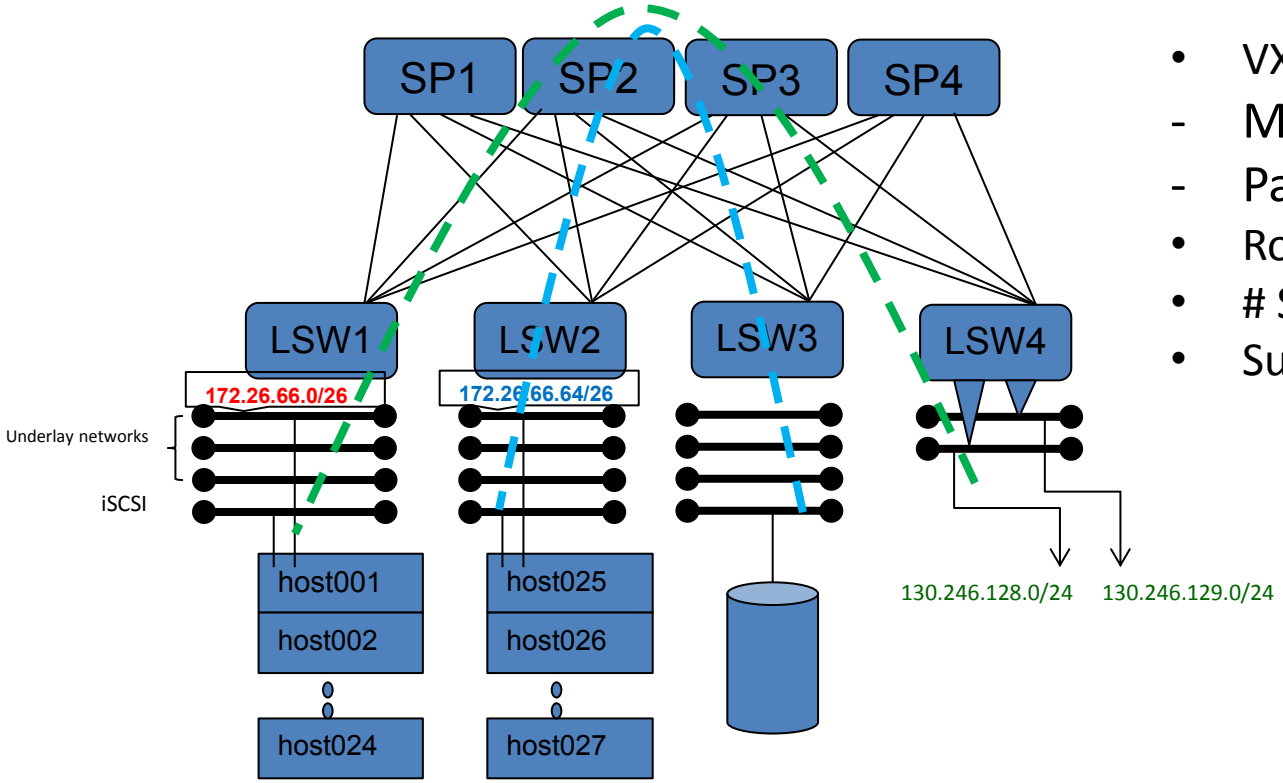
```
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]# ping mgmt.jc.rl.ac.uk
PING mgmt.jc.rl.ac.uk (172.16.151.121) 56(84) bytes of data.
64 bytes from mgmt.jc.rl.ac.uk (172.16.151.121): icmp_seq=1 ttl=60
64 bytes from mgmt.jc.rl.ac.uk (172.16.151.121): icmp_seq=2 ttl=60
64 bytes from mgmt.jc.rl.ac.uk (172.16.151.121): icmp_seq=3 ttl=60
^C
--- mgmt.jc.rl.ac.uk ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2142ms
rtt min/avg/max/mdev = 0.067/0.138/0.230/0.068 ms
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@host052 ~]#
[root@h                ]# traceroute -I mgmt.jc.rl.ac.uk
trac                   ic.rl.ac.uk (172.16.151.121), 30 hops max, 60 byte packets
 1                     26.64.1)  0.437 ms  0.468 ms  0.485 ms
 2                     18.109.5)  0.401 ms  0.406 ms  0.417 ms
 3                     .18.103.66)  0.387 ms  0.409 ms  0.425 ms
 4                     .18.224.1)  0.248 ms  0.381 ms  0.517 ms
 5                     (172.16.151.121)  0.107 ms  0.115 ms  0.115 ms
[root@
```

Fast

Four routed ECMP hops

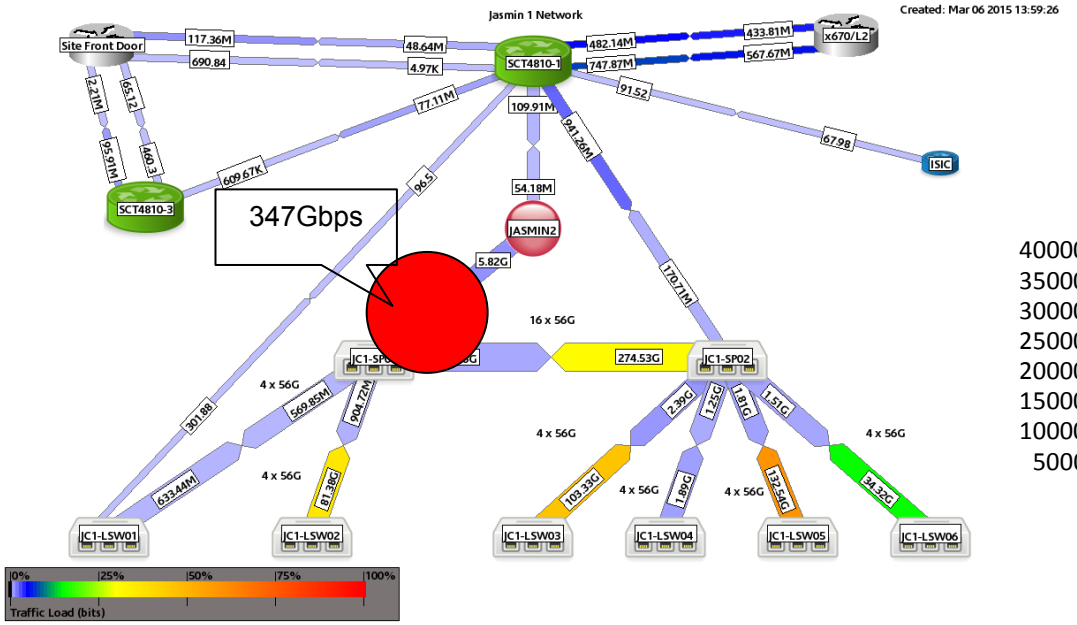


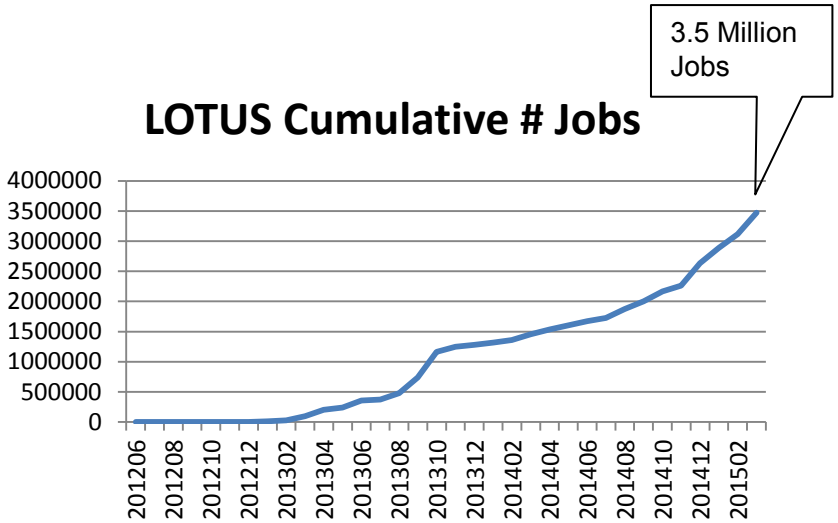
Implications of L3 CLOS for Virtualisation



- VXLAN Overlays
- Multicast routing → PIM
- Panasas access still direct
- Routed iSCSI
- # Subnets
- Sub-optimal IP use

The need for speed





Further info

- JASMIN
  - http://www.jasmin.ac.uk
- Centre for Environmental Data Archival
  - http://www.ceda.ac.uk
- JASMIN paper

  Lawrence, B.N. , V.L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard, and A. Stephens. **Storing and manipulating environmental big data with JASMIN.** *Proceedings of IEEE Big Data 2013, p68-75,* doi:10.1109/BigData.2013.6691556